NIGMS Protein Structure Initiative Meeting Summary
April 24, 1998

Introduction

A one-day meeting, sponsored by the National Institute of General Medical Sciences (NIGMS) at NIH, was held on April 24, 1998, to discuss issues related to a genome-directed Protein Structure Initiative (PSI). The meeting was a follow-up to a larger one held in January 1998 at Argonne National Laboratory. The Argonne meeting surveyed and discussed the proposal that it is now possible to identify, through comparative protein sequence analysis, all protein structural families and then to experimentally determine the 3-D structures of

representative protein molecules from these families. One consequence of such a project would be the discovery of all protein structural motifs and their associated amino acid sequences, which could be used to predict the structural and functional properties of other proteins. To discover all protein folding motifs, it is estimated that on the order of 3,000-5,000 new protein structures must be determined experimentally. Support for this idea was quite widespread, and the purpose of the NIGMS meeting was to provide further discussion of the issues, especially as they might be relevant for a program announcement for individual and program project research grants that bear on this initiative. Below is a summary of the topics discussed at the meeting, with recommendations for further activities.

Background

The Human Genome Project is expected to produce the detailed sequence of the human genome, consisting of over 3 billion bases, by the year 2005. The knowledge of this DNA sequence, which is estimated to code for about 100,000 proteins, is necessary but insufficient for a complete understanding of human and other living systems. The next logical step is to determine the biochemical functions of these 100,000 proteins and their relationships to one another. This task will be significantly enhanced as we deduce the three-dimensional atomic structures of gene products. To determine all 100,000 protein structures is a formidable task, which is out of reach in the foreseeable future. The parsing of the genome into obvious multi-gene families, at about the 25-35 percent sequence identity level, will be a

straightforward task that should reduce this number by an order of magnitude, leaving about 10,000 protein families. Proteins of similar sequence generally share the same structure, so the determination of the structure of one member of a family provides a model for all of them. In favorable cases, knowing the structure may also provide considerable insight into the possible function of a protein, when the function is not already clearly identified from sequence comparisons to proteins with known functions.

Using more sensitive sequence analysis techniques, such as sequence profiling and/or threading methods, it may be possible to reduce the number of distinct families by a factor of 2-3. This should increase the information content per protein structure determined, because the structures of proteins from these families should have a high probability of having either a unique folding motif or being a member of a new structural superfamily. Folding motifs are independently folding units or particular structures that recur in many molecules, families are groups of proteins that show recognizable sequence homology, and superfamilies consist of proteins with similar folding motifs but no discernable sequence similarity. The determination of at least one structure from each of the putative superfamilies suggested by protein sequence analysis should populate the universe of protein folds in a systematic fashion and delineate the set of sequences associated with each folding motif. Many of these sequences will not show any discernable homology to one another, but their identification and assignment to particular folds is essential information that is necessary, for example, to predict protein structure using homology modeling methods.

About 350 folds and 1,200 superfamilies have already been determined experimentally, and projections from these data suggest that there may be on the order of 1,000 folds and 3,000-5,000 superfamilies. As a minimum, to fill in the remaining vacancies in the existing table of protein folds, a PSI will need to determine at least 2,000 new protein structures. However, the practical use of such a table of motifs to predict the structures of other proteins by homology may require multiple examples from each family. Estimates deduced from experience with current homology modeling methods, using examples from known protein families, suggest that the structures of three to five members from such families will be necessary before a practical application of structure prediction techniques can be used in a general fashion. However, if these estimates are reasonable, the initial step of determining, in a concerted and systematic fashion, representative structures for the estimated 1,000 folds from the 3,000-5,000 putative superfamilies identifiable from sequence analysis seems to be a feasible near-term goal.

To determine the structure of at least one member from selected protein families, specific tasks must be carried out. These include:

1. **Identification of protein families and target selection.** Families of proteins can be identified by comparing the protein sequences that are derived from completely sequenced genomes, and targets for structure determination can be selected for families with no apparent sequence homology to proteins of known 3-D structure.

2. **Generation of protein for biophysical analysis.** This

will involve the cloning of selected genes from targeted families into plasmids for overexpression in a suitable microorganism. Overexpressed protein would then be purified for use in crystallization trials or isotopically labeled for NMR analysis.

3. **Structure determination of selected targets.** For those selected proteins that have been expressed, purified, and crystallized, X-ray crystallography can be used to determine the structures. Those proteins that do not crystallize and/or are of appropriate size can be structurally characterized using NMR methods.

4. **Structure and sequence analysis.** The structures that are determined in this project, as well as those determined by other research programs, can be analyzed to evaluate their structural similarity to other known protein structures and to determine the evolutionary relationships that are not identifiable from protein sequence analysis alone.

The intended product of the initial stage of a PSI would be the 3-D atomic structure of at least one member from selected protein families, which is likely to represent either a new fold or a structure from an uncharacterized superfamily. Resultant structural data may be organized into a table as powerful and simple as the periodic table of the elements, which will provide: 1) a "basis set" of folds or structural motifs for use in protein folding studies and structure prediction by homology; 2) good starting points for further experiments, using molecular replacement, to determine the detailed structural characteristics of other members of a family; 3) initial structural information on conserved proteins that will aid

functional studies (e.g., site-directed mutagenesis) and eventually, drug design; and 4) new information for studies of molecular evolution to identify nontrivial structure/sequence relationships, which are essential for the identification of distant phylogenetic connections.

A genome-directed PSI has analogous features to the genome-sequencing project. Initially, DNA sequencing was difficult and its use was restricted to particularly interesting control or coding regions. Gradually, it became apparent that it would be informative to get more complete information, and even to sequence regions that were functionally uncharacterized. As sequencing became more routine and economical, the advantages of complete sequencing projects have become generally accepted.

During the past few years, technological advances have made NMR and crystallographic structure determination more routine and economical. Improvements have been made in screening techniques for crystallization conditions, instrumentation for data collection, phasing methods, electron density map interpretation, and model refinement. Corresponding sets of improvements have also been developed for the determination of protein structures using NMR. Such advances in protein crystallography and NMR methodologies are making this kind of project feasible. While the cost of determining a protein structure is much more than that of sequencing an open reading frame (ORF), protein structure is conserved more strongly than sequence through evolution and it will not be necessary to determine the structures of all gene products. Since we know that proteins with detectable sequence homology will have very similar structures, a great deal of information will be obtained by studying one protein from each family.

Identification of Protein Families and Targ t Selection

The parsing of the genome into families of homologous sequences, and the subsequent selection of proteins with a high probability of having novel structural features, will define the size and scope of this project. It will also focus the attention of the structural biology community toward a set of specific proteins whose structures are likely to provide maximum information content for a given effort. There was a general consensus at the workshop that defendable schemes could be devised for both the generation of protein families and target selection from these families, but there was less agreement on the most appropriate criteria and methods to be used for these tasks.

One such scheme has already been implemented (Tatusov et al., *Science,* 24 October 1997), and families of homologous proteins have been determined by comparative protein sequence analysis. This analysis has attractive features in terms of the phylogenetic distribution of family members, since completely sequenced genomes were involved. This scheme has identified about 1,000 clusters of orthologous groups (*COGs*) by analyzing genomes of microorganisms from the three domains of life (eubacteria, archaebacteria, and eukaryotes). Such clusters appear to correspond to "ancient conserved regions," which may constitute a "basis set" of proteins whose structures and functions form the core biochemical elements from which components of more complex organisms could have

evolved. The determination of the structures and functions of these proteins could, therefore, have profound implications for the understanding of all living systems and their evolutionary relationships. This is only one of several schemes that may be devised to partition the genome into families. Others may involve considerations of protein function, structure, or disease-related issues.

Protein targets for structure studies can be selected by a systematic and rational process of eliminating families whose members have detectable similarity to proteins of known three-dimensional structure. It will be essential to use the most sensitive comparison methods for fold recognition possible, so that the remaining families include proteins with novel folds or distinct versions of known folds. The prioritization of these families for structural analysis could involve considerations of identified, predicted, or uncharacterized function; phylogenetic distribution among the domains of life; and feasibility of structure determination. Families that contain proteins of clearly identified function essential for life and that span all domains of living organisms could constitute a higher priority for structure determination. Many of these targeted proteins may already be in some stage of structural analysis. It will be an important aspect of a genome-directed PSI to identify such cases to eliminate them from consideration for structural analysis. However, all structures that are determined from any source, including this PSI, will feed into the long-term goals of the project. Another priority group could include families with orthologous proteins in a variety of species from more than one domain of life, but having only predicted or uncharacterized functions. Due to their phylogenetic distribution, these proteins may be

expected to possess essential functions, and the determination of their structures may lead to hypotheses of their functions that would stimulate further work and provide unexpected new evolutionary insights.

It was generally felt by the participants at the meeting that another workshop to discuss the issues related to parsing the genome into families and the selection of targets for structural determination would be useful. A tentative date of January 1999 was suggested for such a workshop.

Generation of Protein for Biophysical Analysis

The availability of recombinant DNA technology has made it feasible to generate expression systems that can produce large quantities of almost any protein in a form that is suitable for biophysical analysis. Proteins produced in a PSI will probably come from recombinant DNA sources. The necessary steps for protein production include:

1. **Generate protein expression systems.** The generation of a cDNA clone for any particular ORF of interest, and its incorporation into a suitable expression vector, is a straightforward task that can easily be done in a parallel fashion for high-throughput production. This effort can leverage the work of other genome projects by obtaining clones containing any particular ORF, which must be generated during the complete sequencing of any particular organism. The selection of target proteins for structural analysis from completely sequenced

genomes can take advantage of the availability of these cloned genes. However, even if a clone of a particular protein of interest is not readily available, it has now become a routine molecular biology operation to generate a cDNA clone for almost any particular protein from a wide variety of organisms.

2. **Overexpress the protein.** The expression of protein in tens of milligram amounts for crystallography and NMR is more problematic than the generation of expression vectors. However, there are several expression systems that have been extensively studied. Some of these include: 1) bacterial (*E. coli*), 2) methylotrophic yeast (*Pichia pastorisis*), 3) viral (baculovirus and vaccinia), 4) cell culture (mammalian and insect), and 5) *in vitro* translation. Although the expression of any particular protein may be idiosyncratic, the availability of these and other expression systems significantly increases the ability to produce large quantities of protein. Since it is hard to predict in advance which system will work best for a particular protein, it will be useful to design compatible expression vectors to exchange inserts for the different expression systems.

3. **Purify the protein.** Once an expression system is overproducing protein, then it must be separated from the other contents of the cell system that was utilized for expression. Highly purified protein is required for X-ray and NMR sample preparation and analysis. In general, specific purification schemes and protocols must be designed for each particular protein. For the high-throughput protein production required for a PSI, a more uniform scheme would be desirable. The use of histidine and protein tags

engineered into expression vectors is one possible solution that would make protein purification common to any system.

There is now considerable experience in both the academic and private sectors in the implementation of all three of these required tasks. Although any particular protein may be problematic to express and purify, the availability of the wide variety of methods for protein production that use recombinant DNA technologies should enable the kinds of high-throughput production of proteins necessary for a PSI. The workshop participants felt that protein production might not be a bottleneck for this project.

Preparation of Protein for Biophysical Analysis

Once large quantities of protein are produced, they must be prepared for biophysical analysis. NMR requires the protein to be isotopically labeled, which can be done at the time of expression by enriching cell growth media with appropriately labeled nutrients of foodstuffs. X-ray crystallography requires the generation of protein crystals for diffraction experiments. In the past, growing crystals was a major limitation in the utilization of X-ray crystallography for the determination of the 3-D atomic structures of proteins. In recent years, significant progress has been made in the understanding of the basic mechanisms of crystallization using physical approaches, such as elastic light scattering, interferometry, atomic force microscopy, fluorescence polarization, osmometry, and others. Additionally, practical experience gained mainly by trial and error has produced a set of systematic

procedures, or crystal screens, that have proven to be very effective for the crystallization of a wide assortment of proteins. Some of the factors that are responsible for recent advances and success in growing new protein crystals include: 1) recombinant DNA technology, which enables the production of greater amounts of protein that are easier to purify and are genetically homogeneous; 2) many more investigators are trying to grow crystals for their particular research problems; 3) new reagents and techniques are being devised; and 4) reduced requirements for size, number, and quality of crystals, due to the availability of synchrotron sources, new detectors, and cryocrystallography.

Further work in the systematic study of the chemical and physical properties associated with crystallization should provide additional improvements in the ability to grow protein crystals. Specifically funded projects for technology development in this area would have a positive impact on any protein structure initiative, which requires high-throughput crystal growth in addition to the ability to make large quantities of pure protein. Although additional research in the area of crystal growth is necessary, the attendees at the workshop generally agreed that protein crystal growth would not be a bottleneck in the initial phase of this PSI, where mostly soluble globular proteins would be investigated. However, at least 20 percent of gene products are expected to be membrane proteins. The crystallization of integral membrane proteins is still very problematic. Ultimately, this problem must be solved, but there will be hundreds of nonmembrane proteins whose structures can be determined first. The strategy for this PSI could be to do the "easy ones" first, and then return to the more

difficult cases in later stages of the project.

Structure Determination of Selected Targets

X-ray crystallography and NMR spectroscopy will be the methods used to solve protein structures in a PSI. Although both methods will make major contributions to the initiative, advances in instrumentation, computer technology, and crystal growth place X-ray crystallography as the primary choice for structure determination in a PSI. NMR's primary limitation for structure determination is the size of proteins, which currently is about 30 kD. Newer methods and spectrometers with higher field strength show promise to increase this size limit up to 60 kD in favorable cases, but larger proteins will be problematic. NMR also requires several weeks for data collection and processing, whereas crystallography, in the very best cases, can collect data on a synchrotron in less than 1 hour, with phasing information, and construct a 3-D model in less than 1 week.

There have been significant advances in recent years in the development and implementation of methods and instrumentation for macromolecular crystallography. Access to synchrotron sources, such as the Advanced Photon Source at Argonne National Laboratory, together with CCD detectors and cryocrystallography is making data collection very efficient. It is now possible to collect complete data sets for several proteins in a single day. Using synchrotron sources, crystals as small as 60 μm can be used to solve protein structures. This significantly increases the number of protein structures that may be

determined, because many proteins form only small crystals. Multiple-wavelength Anomalous Diffraction (MAD) techniques coupled with selenomethionine enrichment show promise to nearly eliminate problems associated with phasing. Many of the proteins in a PSI are expected to be amenable to these advanced techniques for data collection and processing. In a project designed for high-throughput structure determination, protein production, crystal growth, and X-ray data collection may not be rate-limiting steps. Given sufficient numbers of crystals, it is now feasible to collect data from several hundred crystals per year, using a dedicated beam line on a third-generation synchrotron. The most time-consuming step is, therefore, likely to be electron density map interpretation. Methods to automate model building and refinement from electron density maps are in the early stages of development, and additional support for research in this area could significantly enhance high-throughput structure determination.

Structure determination using NMR methods has now become routine for proteins less than 30 kD. Using isotopically labeled proteins, resonance assignment has become a straightforward procedure, and methods now exist to automate the interpretation of data and the building of 3-D atomic models. The rate-limiting step has become the collection of data. A 600 MHz spectrometer can collect data for about 10 NMR solution structures per year. It would, therefore, be necessary to have 20 to 40 such machines operating full-time to produce a similar throughput for NMR structure determination as one can obtain at a single beam line on a synchrotron. A "farm" of NMR machines would be necessary for NMR methods to have a significant impact on this PSI. New methods using spectrometers in the 800, 900, and 1,000 MHz range

show promise to increase the minimum size limitation to about 60 kD. However, for NMR to have a significant impact on this PSI, it may be necessary to identify domain size regions in targeted proteins, clone and express these regions, and then solve these structures, which should be a size amenable to NMR methods (200-300 residues). The identification of domains in proteins, primarily from protein sequence analysis, is in its infancy, and advances will be required in such techniques for NMR structure determination to play a major role in high-throughput structural projects. Additionally, subunits and/or domains of larger proteins may not be stable, so that using separated domains may not be a viable strategy in many cases.

Summary and Recommendations

The Human Genome Project has produced complete sequences of several microorganisms, and it is expected to generate the detailed sequences of human and other organisms from the three domains of life by the year 2005. Using computational molecular biology methods, comparative analysis of protein sequences deduced from the DNA sequences will be able to parse these genomes into distinct superfamilies of homologous proteins. Members of these superfamilies all have similar 3-D structures. It is difficult to provide a definitive estimate of the total number of superfamilies with unique folding motifs and/or ones with similar motifs but distinct sequences until more sequences have been determined from complex multicellular organisms. However, reasonable estimates suggest that there may be as few as 3,000 and no more than 10,000 superfamilies with

unique structural and functional properties. The determination of the structure of at least one member from each of these families will populate the space of all protein folding motifs and associated amino acid sequences, which could be used to predict the structures of other proteins. Although a project to determine the structures of several thousand proteins suggested by protein sequence analysis would seem to be formidable, advances in molecular biology and biophysical methods are making the tasks of high-throughput protein production, crystal growth, and X-ray/NMR data collection and analysis a tractable endeavor. It should be possible to determine several thousand novel structures of targeted proteins in a concerted large-scale effort over a period of 5-10 years. The resultant data and associated infrastructure necessary to support such a project will significantly enhance our understanding of genomes, and transform all future structure/function studies.

The specific recommendations from the meeting were to:

1. **Organize a workshop for theoreticians to discuss target selection issues.** The selection of targets using protein sequence analysis will define the size and scope of a genome-directed Protein Structure Initiative. There was general agreement among the participants that defendable strategies for the parsing of the genome into families of homologous proteins and the selection of protein targets with no discernable sequence similarity to known protein structures are feasible. However, there was little consensus as to the most appropriate methods and procedures to be utilized. A proposal was made to hold a workshop, tentatively scheduled for January 1999, to present specific schemes related to issues

of the generation of families of proteins for target selection.

2. **Provide support for the development of necessary technologies and infrastructure.** Significant progress has been made in recent years with respect to the different experimental methods necessary for the implementation of a genome-directed Protein Structure Initiative. Although the specific tasks related to protein production, crystal growth, and X-ray/NMR structure determination have become more sophisticated, further research and development in these areas would enhance the initiative. In particular, it will be necessary to investigate strategies for the integration of these different technologies into a concerted program for high-throughput structure determination. In addition, the establishment of centralized facilities, analogous to the Human Genome Project's sequencing centers, to facilitate the high-throughput production of proteins as well as crystal growth and data collection was considered to be a useful long-term infrastructure objective. Such "protein structure factories" could provide the necessary raw X-ray and NMR data that would subsequently be processed for structure determination in academic and government structural biology laboratories distributed across the country. In this way, the determination of the hundreds of structures targeted by the project could be accomplished most efficiently.

3. **Establish pilot projects in genome-directed structural biology.** It was proposed that several program projects be supported to evaluate different strategies associated with a genome-directed Protein

Structure Initiative. At this time, it is not clear what the best approach may be for an integrated program that includes target selection, protein production, and structure determination. The establishment of several program projects utilizing different overall approaches, but including all the required tasks, would provide the tests from which the necessary experience could be obtained to determine the best strategies for a subsequent large-scale effort.

Paul Bash, Northwestern University (co-chair)
Eaton Lattman, Johns Hopkins University (co-chair)

Workshop Participants

Paul Bash, Northwestern University, Chicago, IL (co-chair)
Eaton Lattman, Johns Hopkins University, Baltimore, MD (co-chair)
Lorena Beese, Duke University, Durham, NC
David Clayton, Howard Hughes Medical Institute, Chevy Chase, MD
David Davies, NIDDK, National Institutes of Health, Bethesda, MD
David Eisenberg, University of California, Los Angeles
Paul Godowski, Genentech Corporation, South San Francisco, CA
Wayne Hendrickson, Columbia University, New York, NY
Sung-Hou Kim, University of California, Berkeley
Eugene Koonin, NCBI, National Institutes of Health, Bethesda, MD
Anthony Kossiakoff, Genentech Corporation, South San Francisco, CA

Alexander McPherson, University of California, Irvine
Gaetano Montelione, Rutgers University, Piscataway, NJ
John Moult, University of Maryland, Rockville
Manuel Navia, Althexis Company, Bedford, MA
Cynthia Peterson, University of Tennessee, Knoxville
Andrej Sali, Rockefeller University, New York, NY
Chris Sander, Millenium Information, Cambridge, MA
Janet Thornton, University College, London, UK
Peter Wolynes, University of Illinois, Urbana-Champaign

Agency Staff Attending

Department of Energy, Germantown, MD
Charles Edmonds
John Wooley

NASA, Huntsville, AL
James Patton Downey
Craig Kundrot

National Science Foundation, Arlington, VA
Lee Makowski
Gerald Selzer
Kamal Shukla

CSR/NIH, Bethesda, MD
Marjam Behar
Nancy Lamontagne
Elliot Postow
Don Schneider
Jean Sipe

NCRR/NIH, Bethesda, MD

Dov Jaron
Marjorie Tingle

NHGRI/NIH, Bethesda, MD
Elise Feingold
Adam Felsenfeld
Elke Jordan

NIGMS/NIH, Bethesda, MD
Jim Cassatt
Marvin Cassman
Jean Chin
Jim Deatherage
Irene Glowinski
Warren Jones
Cathy Lewis
Alisa Machalek
John Norvell
Pat Pillsbury
Martha Pine
Peter Preusch
Marc Rhoades
Sue Shafer
Bert Shapiro
Linda Shein
Helen Sunshine
Janna Wehrle

National Institute of General Medical Sciences
National Institutes of Health
Bethesda, Maryland 20892-6200

Last updated: June 2, 1998